Advanced Algorithms
2nd Project - Approximate Counting

Margarida Silva Martins

***Abstract* –This report presents the work done on the second project of the Advanced Algorithms course.**
The paper will analyse and compare three different ways of counting the number of letters in a text: exact count, approximated count with fixed probability and approximated count with decreasing probability.

I. Introduction

Several areas require counting large number of events.

Counters can be exact in the sense that they count every event or approximated meaning that they count the event with a certain probability.

Approximated counters allow to count a bigger number of events while occupying less space. This approach is useful when analysing large data streams.

In this case the data are text files and the goal is to count the number of occurrences of letters.

The computational efficiency and limitations of three different counting approaches will be analyzed: a exact counter, an approximate counter with fixed probability and an approximated counter with decreasing probability.

II. Data

The texts files used is this work are multiple editions of the literary novel "Don Quixote" by Miguel de Cervantes [1]. There are 5 different files corresponding to the novel written 5 different languages: spanish, english, dutch, hungarian and french. All this files were downloaded from the Project Gutenberg [2].

The Project Gutenberg file headers were removed and the letters converted to capital letters because it is not relevant if the letter is capital or lowercase.

In order to compare the frequency of letters between languages the counters ignore accents, for example an Â is converted to A.

III. Counters Algorithms

A. Exact counter

The exact counter algorithm loops over each text file letter and increments the respective letter counter by one.

This means that with n letters the sum of all letter counters will be n which needs $log_2(n)$ bits in order to represent it.

B. Approximated counter with fixed probability

The approximated counter with fixed probability loops over each text file letter and increments the respective letter counter with probability 1/64.

This means that with n letters the sum of all letter counters is expected to be approximately $n/64$ which needs $log_2(n/64) = log_2(n) - 6$ bits in order to represent it which is 6 bits less than the exact counter.

C. Approximated counter with decreasing probability

The approximated counter with decreasing probability loops over each text file letter and increments the respective letter counter with probability $1/(\sqrt{2})^k$, with k being the current counter value.

This means that while the previous two counters were linear this is a logarithmic counter. As the counter value increases the probability of counting decreases.

With n letters the sum of all letter counters is expected to be $floor(log_{\sqrt{2}}(n+1)) = floor(2log_2(n+1))$ which needs $log_2(2log_2(n+1)$ bits in order to represent it. For large amounts of data this is significantly lower than the previous two approaches.

IV. Results and Counters Comparation

A. Tests Done

For each text file the exact letters count was calculated once while the approximated counters were repeated 20 times for each type of counter. In the exact counting, for each letter, it was recorded the absolute counter value and the relative counter value.

In the approximated counting it was recorded, for each letter, the average absolute estimated count, the average relative estimated count, the highest estimated value and the lowest estimated value as well as the average absolute counter value, the average relative counter value, the highest counter value and the lowest counter value.

B. Exact and Estimate Count Values

As it can be observed in figure 1, the estimations made by the fixed probability counter are much closer to the ones made by the decreasing probability counter.

It is important to notice that in some cases such as the A and D letters the fixed probability counter estimates an higher occurrence value than the real one, this does not happen with the decreasing probability counter.

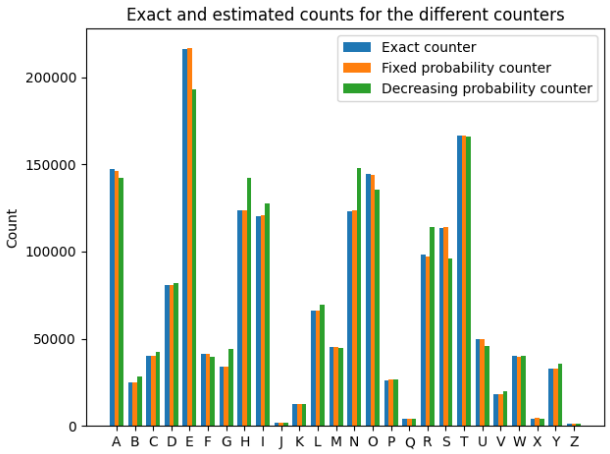

Fig. 1 - Comparison of the exact or estimated average values for the three different counters using the english file.

Figure 2 shows the relative exact or estimated counts using the same file as the previous figure. The most frequent letter E amounts 12% of the total letter occurrences with the second most frequent letter being bellow 10%.

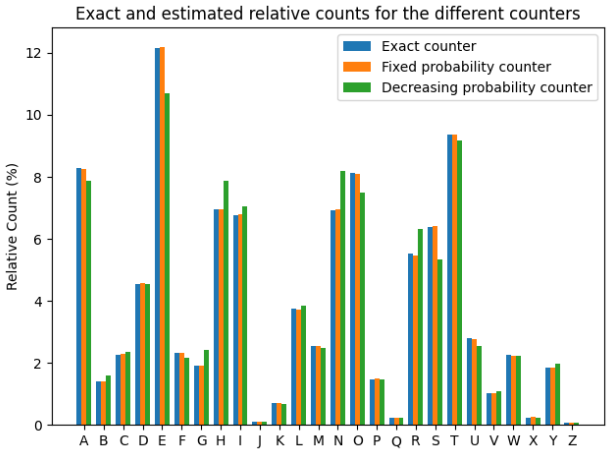

Fig. 2 - Comparison of the exact or estimated average relative values for the three different counters using the english file.

Using the relative frequency of letters we can observe whether the approximate counts identify or not the same most frequent letters as shown in table I.

Fixed probability counter lists the same ten most frequent letters as the exact counter. This is not true for the decreasing probability counter that puts the 6^{th} most frequent letter N in 3^{rd} place. It also mistakes the positions of A, O, S and R letters.

letter	Position EC	Position FPC	Position DPC
E	1	1	1
T	2	2	2
A	3	3	4
O	4	4	6
H	5	5	5
N	6	6	3
I	7	7	7
S	8	8	9
R	9	9	8
D	10	10	10

TABLE I

Top 10 most frequent letters using relative values for the exact counter (EC), fixed probability counter (FPC) and decreasing probability counter (DPC) using the english text

C. Estimated count errors

C.1 Fixed probability counter

Table II presents the mean absolute and relative errors for the fixed probability counter for the spanish text file.

Until the 12^{th} most frequent letter the relative error is less than 1%. As the letters become less frequent the relative error tends to increase but with the exceptional case of the letter W it is always less than 4%.

letter	EC	FPC	MAE	MRE (%)
E	230608	230070	538	0,23
A	201789	201098	691	0,34
O	163637	164166	109	0,07
S	126534	126643	109	0,09
N	113486	113216	270	0,24
R	101531	102019	488	0,48
I	90581	90890	309	0,34
L	89695	89334	361	0,40
D	87941	88124	183	0,21
U	80147	80400	253	0,32
T	62246	61808	438	0,70
C	59914	60595	681	1,14
M	44836	44566	270	0,60
P	35600	36051	451	1,27
Q	32708	33178	470	1,44
Y	25213	25194	19	0,07
B	24277	24493	216	0,89
H	20009	19917	92	0,46
V	17993	17939	54	0,30
G	17345	17574	229	1,32
J	10615	10291	324	3,04
F	7624	7507	117	1,53
Z	6543	6448	95	1,45
X	380	371	9	2,39
W	2	0	2	100,00

TABLE II

Mean absolute and relative errors (MAE and MRE) for fixed probability counter (FPC) for spanish text file

In Table III below the lowest and highest frequency values as well as the maximum absolute and relative errors are shown for the fixed probability counter for the spanish text file.

This error values are significantly bigger than the mean ones, for example, in the most frequent letter, E, the maximum relative error is almost 15 times higher. This shows the importance of multiple tests.

Just like the mean error higher letter frequency values correspond to lower maximum relative errors, with the two most frequent letters A and E having relative errors below 4%.

letter	EC	FPC HV	FPC LV	MAE	MRE (%)
E	230608	233984	222976	7632	3,42
A	201789	206336	195200	6589	3,27
O	163637	173440	159232	9803	5,99
S	126534	132800	122240	6266	4,95
N	113486	118272	108032	5454	4,81
R	101531	106688	96128	5403	5,32
I	90581	94528	87104	3947	4,36
L	89695	94784	84416	5279	5,89
D	87941	92672	83072	4869	5,54
U	80147	86848	75008	6701	8,36
T	62246	66304	59136	4058	6,52
C	59914	65024	57408	5110	8,53
M	44836	50368	40896	5532	12,34
P	35600	39360	34048	3760	10,56
Q	32708	36864	30464	4156	12,7
Y	25213	27200	23360	1987	7,88
B	24277	27392	21248	3115	12,83
H	20009	22272	18368	2263	11,31
V	17993	19584	15872	2121	11,79
G	17345	19520	15872	2175	12,54
J	10615	11712	9280	1335	12,58
F	7624	8576	6336	1288	16,89
Z	6543	7552	4800	1743	26,64
X	380	704	192	324	85,26
W	2	0	2	2	100,00

TABLE III

Maximum absolute and relative errors (MAE and MRE) for fixed probability counter (FPC) for spanish text file

C.2 Decreasing probability counter

Table IV presents the mean absolute and relative errors for the decreasing probability counter for the spanish text file.

Although some letters present small relative errors, letter I with 2,56% or letter Q with 0,32%. Most relative error values are above 5% which is considerably higher compared with the fixed probability counter.

Contrary to the fixed probability counter results higher frequency letters do not mean lower relative error, for example letter N is the 5^{th} most frequent letter and has the 5^{th} biggest relative error value.

letter	EC	DPC	MAE	MRE (%)
E	230608	210481	20127	8,72
A	201789	183472	18317	9,08
O	163637	149991	13646	8,34
S	126534	142279	15745	12,44
N	113486	98289	15197	13,39
R	101531	109477	7946	7,83
I	90581	88260	2321	2,56
L	89695	85322	4373	4,88
D	87941	93713	5772	6,16
U	80147	86621	6474	8,08
T	62246	59781	2465	3,96
C	59914	66943	7029	11,73
M	44836	47725	2889	6,44
P	35600	39284	3684	10,35
Q	32708	32602	106	0,32
Y	25213	22474	2739	10,86
B	24277	27673	3396	13,99
H	20009	19302	707	3,66
V	17993	16735	1258	6,99
G	17345	18713	1368	7,89
J	10615	11484	869	8,19
F	7624	7203	421	5,52
Z	6543	6061	482	7,37
X	380	322	58	15,26
W	2	3.05	1,05	52,5

TABLE IV

Mean absolute and relative errors (MAE and MRE) for decreasing probability counter for spanish text file

In Table V below the lowest and highest frequency values as well as the maximum absolute and relative errors are shown for the decreasing probability counter for the spanish text file.

Similary to the fixed probability values this error values are significantly bigger than the mean ones. There are only two letters, J and F, that have a relative error below 50%. On the other hand 7 letters have an relative error bigger than 100% with letter R reaching 340%. This means that without multiple tests this counting approach is not accurate.

letter	EC	DPC HV	DPC LV	MAE	MRE (%)
E	230608	316435	111876	118732	51,49
A	201789	316435	79108	122681	60,80
O	163637	223753	55937	107700	65,82
S	126534	316435	79108	189901	150,08
N	113486	223753	39553	110267	97,16
R	101531	447507	55937	345976	340,76
I	90581	158217	55937	67636	74,67
L	89695	158217	39553	68522	76,39
D	87941	158217	39553	70276	79,91
U	80147	223753	39553	143606	179,18
T	62246	158217	19776	95971	60,66
C	59914	158217	27968	98303	62,13
M	44836	111876	13984	67040	59,92
P	35600	79108	19776	43508	122,21
Q	32708	55937	19776	23229	71,02
Y	25213	39553	9888	15325	60,78
B	24277	55937	9888	31660	130,41
H	20009	39553	9888	19544	97,68
V	17993	39553	6991	21560	119,82
G	17345	39553	6991	22208	123,44
J	10615	19776	4943	9161	46,32
F	7624	9888	4943	2681	35,17
Z	6543	9888	2471	4072	62,23
X	380	617	154	237	62,68
W	2	4	2	2	100,00

TABLE V

Mean absolute and relative errors (MAE and MRE) for decreasing probability counter (DPC) for spanish text file

D. Counters size

Table VI bellow represents the counter value for the three different counting approaches and the number of bits necessary to represent it using the spanish text file.

As expected the exact counter requires more bits than the approximated approaches with the decreasing approach requiring the least.

Using the formulas in chapter III if the exact counter requires n bits the fixed probability counter would need n-6 bits. This is correct for all the letters with the exception of letter Q where the difference is 5.

All the counts using the decrease probability counter are under 1 byte. In total the decrease probability counter needs around on third of the space occupied by the exact counter. If we would only count the total number of letters this difference would be lower in relative terms. The exact counter would need 22 bits while the decrease probability counter 11 bits 50% less.

It is also important to notice that in all the 20 times the fixed probability approach did not detect the lowest frequency letter, W.

letter	EC	nbits	FPC	nbits	DPC	nbits
E	230608	19	3611	13	33	7
A	201789	19	3128	13	33	7
O	163637	19	2552	13	32	6
S	126534	18	1976	12	31	6
N	113486	18	1776	12	31	6
R	101531	18	1570	12	31	6
I	90581	18	1432	12	31	6
L	89695	18	1407	12	30	6
D	87941	18	1369	12	30	6
U	80147	18	1244	12	30	6
T	62246	17	963	11	29	6
C	59914	17	933	11	29	6
M	44836	17	705	11	28	6
P	35600	17	550	11	27	6
Q	32708	16	519	11	28	6
Y	25213	16	394	10	27	6
B	24277	16	382	10	27	6
H	20009	16	312	10	25	6
V	17993	16	280	10	25	6
G	17345	16	277	10	26	6
J	10615	15	165	9	23	6
F	7624	14	121	8	23	6
Z	6543	14	105	8	22	6
X	380	10	6	4	14	5
W	2	2	0	0	2	2
Total	1651254	402	25776	258	640	147

TABLE VI

COUNTER VALUES FOR EXACT COUNTER (EC), FIXED PROBABILITY COUNTER (FPC) AND DECREASING PROBABILITY COUNTER (DPC) USING THE SPANISH FILE AND RESPECTIVE MINIMUM REQUIRED NUMBER OF BITS

E. Language letter occurencies

Figures 3, 4 and 5 compare the relative letter counts for the 5 different languages using the three different counter approaches.

With the exception of hungarian, the letter E is the most used reaching values above 17% for the french and dutch languages. However only spanish and french share the same second most used letter, A. Some languages have really low occurences for certain letters (¡0.01%) as is the case of hungarian, spanish and french for the W letter.

The relative values for the fixed probability counter are very similar to the exact ones. However in the decreased counter they vary more. For example french appears as the language with biggest relative occurences of the letter E opposed to ducth. This also happens with letters T and S where the english and hungarian places are switched.

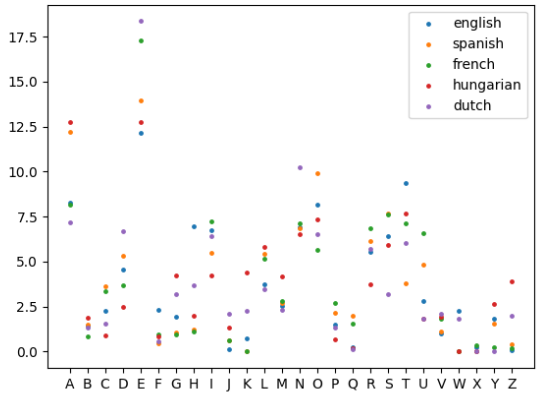

Fig. 3 - Relative frenquency of letters using the exact counter for 5 different languages.

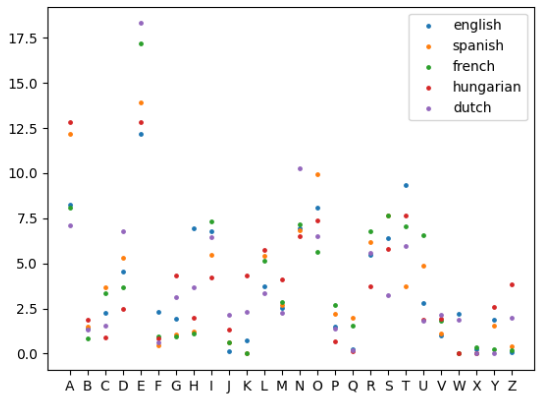

Fig. 4 - Relative frenquency of letters using the fixed probability counter for 5 different languages.

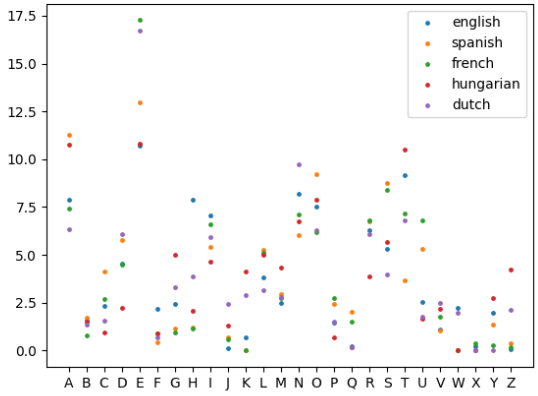

Fig. 5 - Relative frenquency of letters using the decreasing probability counter for 5 different languages

V. Conclusion

With this work it can be concluded that with medium to large values approximated counters with fixed probability can provide an accurate count of events occupying slightly less space than exact counters.

Approximated counters with decreasing probability give less accurate results but for large number of events they occupy much less space. With a large amount of data where its not needed to know the exact number of events but the order of magnitude this type of counters can be really useful. However counting letter occurrences even with larger literary works does not have the size required in order to make the use of approximated counters with decreasing probability a better option.

References

[1] "Dom quixote".
URL: `https://en.wikipedia.org/wiki/Don_Quixote`

[2] Project Gutenberg, "Books by cervantes saavedra, miguel de".
URL: `https://www.gutenberg.org/ebooks/author/505`